Data Mining

ID2222

Homework 2: Discovery of Frequent Itemsets and Association Rules

Authors:

Abdul Aziz Alkathiri

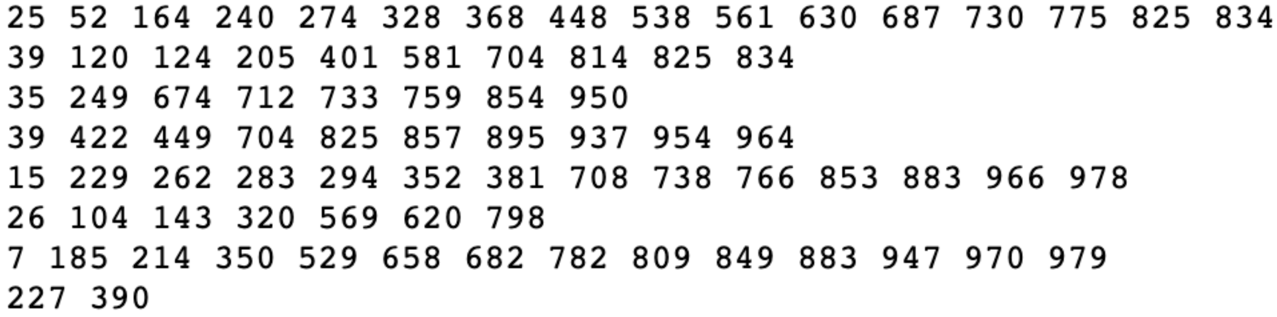
Francesco Staccone

November 17, 2019

1 Introduction

The dataset we used is the recommended one: T10I4D100K.dat.

The dataset contains transactions of items from retail market basket data. The dataset as used contains one column which contains the items bought within one transaction, as shown in Figure1. More about the dataset can be read here.



```
25 52 164 240 274 328 368 448 538 561 630 687 730 775 825 834
39 120 124 205 401 581 704 814 825 834
35 249 674 712 733 759 854 950
39 422 449 704 825 857 895 937 954 964
15 229 262 283 294 352 381 708 738 766 853 883 966 978
26 104 143 320 569 620 798
7 185 214 350 529 658 682 782 809 849 883 947 970 979
227 390
```

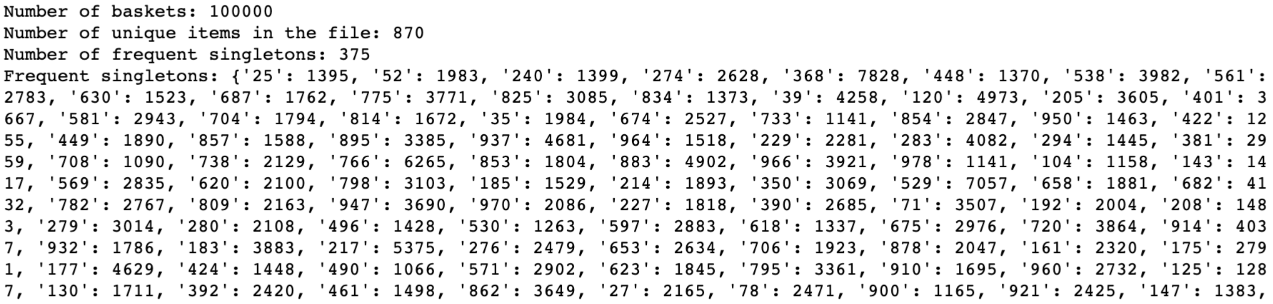
Figure 1: The retail basket dataset.

Our submission consists of one .ipynb file that implements the frequent itemsets and association rules, and the dataset itself.

2 Implementation

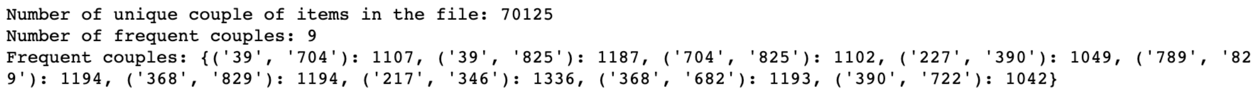
2.1 Frequent itemsets

We used some parts of itertools and collections for combinations and dictionaries. The apriori function we implemented, considered the support threshold selected (1% of all the baskets) is composed of three main steps. The first one selects the singletons, the second one selects the couples and the third one selects the triplets. Since only one triplet was found according to the selected support threshold, there was no need to look for bigger tuples.
The whole algorithm leverages on the idea that while iterating on each basket, it generates only the combinations of its items already contained in the previous step itemset, so that is produces only the necessary couples and saves computation. In general: if X is a frequent k-item set, then all (k-1)-item subsets of X must also be frequent, that's why it works.
Results are visible in Figure 2, 3, 4.



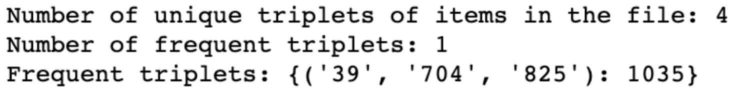
```
Number of baskets: 100000
Number of unique items in the file: 870
Number of frequent singletons: 375
Frequent singletons: {'25': 1395, '52': 1983, '240': 1399, '274': 2628, '368': 7828, '448': 1370, '538': 3982, '561':
2783, '630': 1523, '687': 1762, '775': 3771, '825': 3085, '834': 1373, '39': 4258, '120': 4973, '205': 3605, '401': 3
667, '581': 2943, '704': 1794, '814': 1672, '35': 1984, '674': 2527, '733': 1141, '854': 2847, '950': 1463, '422': 12
55, '449': 1890, '857': 1588, '895': 3385, '937': 4681, '964': 1518, '229': 2281, '283': 4082, '294': 1445, '381': 29
59, '708': 1090, '738': 2129, '766': 6265, '853': 1804, '883': 4902, '966': 3921, '978': 1141, '104': 1158, '143': 14
17, '569': 2835, '620': 2100, '798': 3103, '185': 1529, '214': 1893, '350': 3069, '529': 7057, '658': 1881, '682': 41
32, '782': 2767, '809': 2163, '947': 3690, '970': 2086, '227': 1818, '390': 2685, '71': 3507, '192': 2004, '208': 148
3, '279': 3014, '280': 2108, '496': 1428, '530': 1263, '597': 2883, '618': 1337, '675': 2976, '720': 3864, '914': 403
7, '932': 1786, '183': 3883, '217': 5375, '276': 2479, '653': 2634, '706': 1923, '878': 2047, '161': 2320, '175': 279
1, '177': 4629, '424': 1448, '490': 1066, '571': 2902, '623': 1845, '795': 3361, '910': 1695, '960': 2732, '125': 128
7, '130': 1711, '392': 2420, '461': 1498, '862': 3649, '27': 2165, '78': 2471, '900': 1165, '921': 2425, '147': 1383,
```

Figure 2: Singletons



```
Number of unique couple of items in the file: 70125
Number of frequent couples: 9
Frequent couples: {('39', '704'): 1107, ('39', '825'): 1187, ('704', '825'): 1102, ('227', '390'): 1049, ('789', '82
9'): 1194, ('368', '829'): 1194, ('217', '346'): 1336, ('368', '682'): 1193, ('390', '722'): 1042}
```

Figure 3: Couples

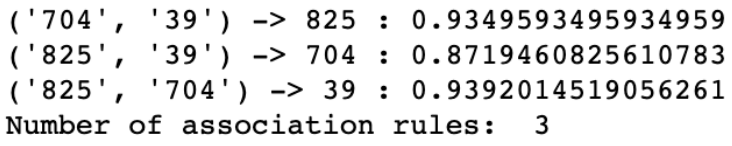


```
Number of unique triplets of items in the file: 4
Number of frequent triplets: 1
Frequent triplets: {('39', '704', '825'): 1035}
```

Figure 4: Triplets

2.2 Association rules

Next we implemented the association rules part, selecting 0.8 as confidence threshold. The algorithm iterates over the tuples sets to find associations, select the support value of the current tuple, creates some subsets of the tuple computing the relative supports and computes the confidence values for the selected association rule. Then the association rules are filtered according to the confidence threshold value. Results are visible in Figure 5.



```
('704', '39') -> 825 : 0.9349593495934959
('825', '39') -> 704 : 0.8719460825610783
('825', '704') -> 39 : 0.9392014519056261
Number of association rules:  3
```

Figure 5: Association Rules